Synthese (2011) 179:223–238
DOI 10.1007/s11229-010-9779-2

Dynamics we can believe in: a view from the Amsterdam School on the centenary of Evert Willem Beth

Cédric Dégremont · Jonathan Zvesper

Received: 11 July 2009 / Accepted: 17 November 2009 / Published online: 9 October 2010

Abstract Logic is breaking out of the confines of the single-agent static paradigm that has been implicit in all formal systems until recent times. We sketch some recent developments that take logic as an account of information-driven interaction. These two features, the dynamic and the social, throw fresh light on many issues within logic and its connections with other areas, such as epistemology and game theory.

Keywords Dynamic logic · Interaction · Information update · Belief change · Epistemology · Game theory

Logic is breaking out of the confines of the single-agent static paradigm that has been implicit in almost all formal logic until recent times. That is the message of van Benthem (1996, 2003). In this paper we sketch some recent developments in this direction, part of a new paradigm that sees logic as an account of information-driven agency, which is typically multi-agent and interactive. These two features, the dynamic and the social, throw fresh light on many issues within logic and its connections with other areas.

Our focus will be on how these themes have been developed within our own research community. However, in keeping with the interactive character of these developments, our aim in this paper is not to be insular! We rather wish to highlight the fruits that can be bourn by reflecting on how dynamics and interaction, concrete or abstract, can be introduced into traditionally static or mono-agent disciplines.

After introducing the key notions of dynamic epistemic logic in Sect. 1, we will see an application, due to van Benthem (2004) in formal epistemology, specifically to

C. Dégremont (✉) · J. Zvesper
The University of Amsterdam, Building C, Room C3.119 I, Science Park 904,
1098 XH Amsterdam, The Netherlands
e-mail: cedric.uva@gmail.com

Fitch's paradox, before summarising some logical work on dynamics of *belief* rather than *knowledge* in Sect. 3. Then we will mention applications in game theory, and focus on such applications in Sects. 5 and 6.

1 From static to dynamic epistemic logic

We do not pretend to have a definition of Logic as a field of inquiry, but it could be characterised approximately as the formal study of reasoning. Traditionally, the particular object of this study has, albeit implicitly, concerned a single agent reasoning in isolation. To put it another way: logicians have studied valid inference without any reference to who is doing the inferring.

Although in many cases we do reason in isolation, there is clearly a natural multi-agent component to much of our actual reasoning.[1] This takes two forms: we reason *about* other people, and also *with* other people.

Hintikka (1962) can be credited with a systematic introduction of the explicit reasoning subject into the domain of logic, with the invention of what is known as 'epistemic logic', studying validities for logical operators K_i and B_i, meaning respectively 'i knows …' and 'i believes …'. (Hintikka does also briefly discuss the multi-agent case, with operators K_a, K_b etc., remarking for example that $K_a K_b \varphi \rightarrow K_a \varphi$ should be a theorem.)

We assume some familiarity with modal languages and their relational semantics; for more details see e.g. Blackburn et al. (2001). Let EL denote the basic modal language, obtained by adding modalities K_i for each i of a fixed set of agents, to some base propositional language. We interpret EL over relational models. Relational semantics is a possible worlds semantics, with a relation $\sim_i$ for each modality K_i. If $s \sim_i t$, then if the actual world were s then i would consider it possible that the world is actually t. In order to interpret the operator $K_i \varphi$ as meaning that i *knows* that φ, it is natural to ask that $\sim_i$ be an equivalence relation, and so K_i be an S5 modality.

The concept of *common knowledge*, first formalised in Aumann (1976), is also often considered in epistemic logic, by including an operator $C_G \varphi$ for the *group* of agents G. Intuitively $C_G \varphi$ means that all agents in G know φ, all agents in G know that all agents in G knows φ and so on. Semantically, this fixpoint notion corresponds to taking the transitive closure of the union of the relations for the agents in G.

Building on this static epistemic logic, 'dynamic epistemic logic' (DEL) includes into the language operators $\langle \alpha \rangle$, meaning '*after the event* α *occurs* …'. In adding these dynamic operators to static epistemic logic, DEL merges ideas from philosophy and computer science, and has been the engine of the research program described in this paper. Baltag et al. (1999) and Gerbrandy (1999) were seminal in the development of DEL. van Ditmarsch (2000) found applications of DEL to reasoning about parlour games, and the first textbook devoted to DEL has recently appeared (van Ditmarsch et al. 2007). DEL is a generalisation of public announcement logic (PAL, Plaza 1989). PAL includes modalities $\langle !\varphi \rangle$, meaning '*after* φ *is (publically and truth-*

[1] Furthermore, if we are to follow Wittgenstein, then language itself and so also (symbolic) reasoning are fundamentally social phenomena.

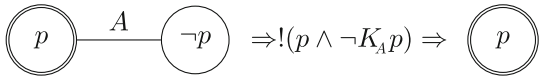


Fig. 1 A public announcement of a Moore-like sentence

fully) announced, …'. Given an epistemic model $\mathcal{M}$, let $\mathcal{M}|\varphi$ be its relativization to $[\![\varphi]\!]^{\mathcal{M}} = \{w \mid \mathcal{M}, w \Vdash \varphi\}$. The truth condition for public announcements is then:

Definition 1 (Truth condition for public announcements)

$$\mathcal{M}, w \Vdash \langle !\varphi \rangle \psi \text{ iff } \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}|\varphi, w \Vdash \psi$$

A pleasing feature of the model-changing PAL/DEL approach is that it correctly handles "blindspots" (Sorensen 1988) like Moore sentences $p \wedge \neg K_i p$. The essential point about these sentences, that is clarified by DEL, is that they can be examples of *true sentences that cannot be learned*. Figure 1 shows what happens when we apply a simple public announcement of such a sentence to a model: Circles are states (possible worlds), and lines labeled with a letter indicate the 'indistiguishability relation' $\sim_i$ for the corresponding player (we do not draw the reflexive lines, nor will we draw the transitive lines in what follows). The actual world is denoted by a double circle. There initially A does not know p, then there is an announcement that "p is true but A does not know it", so an announcement $!\varphi :=!(p \wedge \neg K_A p)$. The important point from this elementary example is just that *after* the announcement, A does know p, and knows that she knows it. Therefore after the announcement of φ, φ is not known! Notice though that this fits with our intuition: what we would learn, if somebody tells us a Moore sentence, is that it *was* true at the moment just before it was uttered.[2]

The dynamics of information are thus well handled by PAL, but PAL only allows for very simple kinds of information. What about other epistemic events than just announcements? Interesting epistemic events that are *not* covered by public announcement logic include lies, partial information, private announcements, …. Gerbrandy and Groeneveld (1997) considered one small generalisation: public announcements to subgroups. Full DEL allows much more. The key idea is to use *models* for events, just as we have models for states in standard (static) epistemic logic.

Definition 2 An event model for the agents N is a tuple $(E, (\sim_i)_{i \in N}, PRE)$, where E is a non-empty set of events, each $\sim_i$ is an equivalence relation over W, and $\mathtt{pre} : E \to \mathcal{L}$ gives the 'precondition' of each event.

In order to combine static models with event models, DEL uses the product operation given in Definition 3. The idea is to interpret the indistinguishability relation in the event models in the same way as in the static models, so that $e \sim_i d$ means that if the event e *actually happens*, then agent i considers it possible that what is actually happening is d. Definition 3 is just a formal working out of this intuition.

[2] "Moore sentences" are named G.E.Moore, who remarked on the paradoxical nature of the first-person statement 'It is raining and *I* don't believe it.' (Moore 1942). The correct treatment of Moore sentences by DEL will be relevant to the discussions below of AGM belief revision and of Fitch's paradox.

Definition 3 Given an epistemic model $\mathcal{M} = (W, (\sim_i)_{i \in N}, V)$ and an event model $\mathcal{E} = (E, (\sim_i)_{i \in N}, \mathtt{pre})$, the product update of $\mathcal{M}$ by $\mathcal{E}$, written $\mathcal{M} \otimes \mathcal{E}$, is the epistemic model with the domain of pairs $\{(w, e) \in W \times E \mid \mathcal{M}, w \Vdash PRE_e\}$, the relation $(w, e) \sim_i (w', e')$ iff $w \sim_i w'$ and $e \sim_i e'$, and the valuation of p is $\{(w, e) \in W \times E \mid w \in V(p)\}$.

The idea is that when you apply an event model to a state model, you generate a new state model, in which the states are just the pairs (s, e), where s was a state in the original model and e is an event that could in principle have occured. Those events that could not have occured are ruled out by the precondition clause: if $s \Vdash \neg PRE_e$, then e could not in principle occur. Then the definition of the new relation in the new model is indeed a straightforward working out of the intuition described above: $(s, e) \sim_i (t, d)$ means that i considers (t, d) possible if the actual state was s and e happened. This will hold only if: (a) at s, i considered t to be possible, and (b) when e occurs, i considers d to be possibly occuring.

It is then easy to see how we might define a language with action modalities of the form $\langle \Sigma, \mathbf{e} \rangle$, where Σ is an action model and e an event in it, with the following semantics:

Definition 4 (Truth definition for epistemic events)

$$\mathcal{M}, w \Vdash \langle \Sigma, \mathbf{e} \rangle \varphi \text{ iff } \mathcal{M}, w \Vdash \mathtt{pre}(e) \text{ and } \mathcal{M} \otimes \Sigma, (w, e) \Vdash \varphi$$

As an illustration of DEL, suppose that three people, Alice Bob and Claire, are playing cards. To make the example very simple, suppose that there are only three cards, 1, 2 and 3, distributed one to each player. Then there are 6 possible scenarios. If we suppose that all this is common knowledge among Alice Bob and Claire, then we can draw the static epistemic model given in Fig. 2. The states are labelled according to the distribution of cards, so that in the actual world, Alice has 1, Bob has 2 and Claire has 3. Now suppose that Bob shows Claire his card. *Part* of this action is effectively an announcement to Bob and Claire, that Claire knows to be truthful, that Bob has card 2. We model this with the event model that is depicted in Fig. 3, where the language has propositional variables of the form in meaning that player i has card n, e.g. $a1$ means Alice has 1. Here the preconditions are written inside the nodes representing the states. In this model, Alice is uncertain what event has actually taken place. That is, Alice knows that Bob has shown Claire his card (we assume that everybody sees everybody seeing Bob's action), but she doesn't know which precondition the action had. That is, for all Alice knows, Bob showed Claire card 3. It might seem odd to the reader that in the event model Alice considers it possible that $b1$ was the precondition for the action. Of course, *in the actual context* Alice will not consider this possibility, and the concerned reader can check that the updated model, as depicted in Fig. 4, does indeed conform with her understanding of what would happen. The point is that we, in effect, encode *Bob*'s uncertainty via Alice's relation, in the sense that Bob considers it possible that Alice has card 3, and so considers it possible that Alice considers it possible that the event that took place was Bob showing Claire 1. In this particular situation, the effect of the event model has in effect been to cut some indistinguishability links from the original model. However, in general the results of

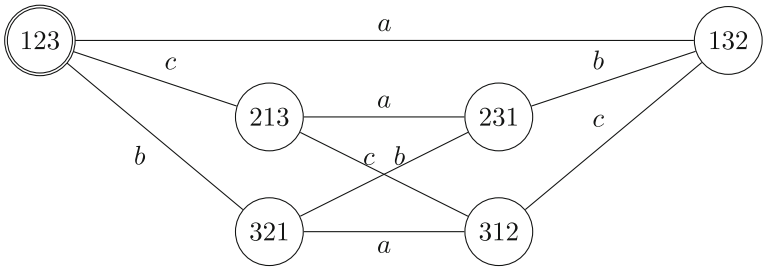
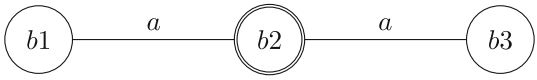


Fig. 2 A model of the initial epistemic situation of Alice, Bob, and Claire

Fig. 3 An event model of Bob showing Claire his card



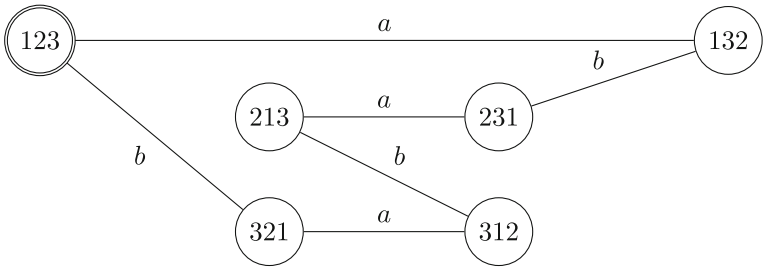


Fig. 4 A model of the situation after Bob has shown Claire his card

event models can be much more drastic, expanding the model or reducing it in many ways.

Indeed, Baltag and Moss (2004) go so far as to state that that DEL is in some sense complete for social epistemic situations. This statement is made in the form of two theses that we do not reproduce here since they are long and carefully worded but the gist is that DEL essentially as it is presented here is able to capture every epistemic feature of any social situation including its dynamics. So the example we gave with cards above would just be an instantiation of this, and every such situation could be represented by a static model and an event model. Their theses are reminiscent of the Church-Turing thesis, in that they relate an informal set of phenomena with a formal mathematical definition. As such it is of course hard if not impossible to verify the thesis. They are bold claims but far from implausible.

What about the logics themselves? How do we axiomatise PAL or DEL? This depends on the underlying language, but if we take it to be the basic modal language EL, things are quite nice for the logician. Indeed by mean of a compositional analysis we can show e.g. that EL is at least as expressive as PAL. A compositional analysis gives us a procedure to recursively translate any formula from PAL+EL back into EL. All we have to do it to check whether each clause or *reduction axiom* is sound.

Proposition 5 (Compositional Analysis of PAL) *The following axioms are sound with respect to the class of all epistemic models.*

1. $\langle !\varphi\rangle p \leftrightarrow (\varphi \wedge p)$
2. $\langle !\varphi\rangle \neg\psi \leftrightarrow (\varphi \wedge \neg\langle !\varphi\rangle\psi)$
3. $\langle !\varphi\rangle(\psi \vee \chi) \leftrightarrow (\langle !\varphi\rangle\psi \vee \langle !\varphi\rangle\chi)$
4. $\langle \varphi K_i \psi \leftrightarrow (\varphi \wedge K_i \rightarrow\rangle\varphi\psi)$

Furthermore, given any set of event models, axiomatizing the dynamic epistemic logic having modalities for all these events can also be done by means of reductions axioms following the following general scheme.

Proposition 6 (Action-Knowledge reduction axiom Baltag et al. 1999) *The following axiom is sound on the class of all epistemic models*

$$[\mathbf{\Sigma},\mathbf{e}]K_i\varphi \leftrightarrow (\mathtt{pre}(e) \rightarrow \bigwedge\{K_i[\mathbf{\Sigma},\mathbf{f}]\varphi \ : \ e \sim_i f\})$$

Logically speaking, PAL internalises relativisation. Therefore some readers might not find it surprising that many formal languages are closed for public announcements, i.e. that adding public announcement operators does not increase the expressivity of the underlying static language. But let us consider the language of epistemic logic with common knowledge. In this case adding public announcement operators strictly increases the expressive power of the logics. In such a case completeness via reduction axioms is no longer an option, the logicians has two options: prove completeness with Henkin-style tools or extend the underlying static language in such a way that it becomes closed again under relativisation (see van Benthem et al. 2006).

2 Formal epistemology

We will look here at one example where the dynamic logic approach has been brought to bear on one topic from formal epistemology: the problem of Fitch's Paradox.

Some forms of anti-realism uphold the *verificationist* thesis, that *every truth can be known*. Letting $\Diamond$ capture what we mean here by '*can*', we formalise the verificationist thesis by the following schema, that we are allowed to instantiate with any formula φ:

$$\varphi \rightarrow \Diamond K_i\varphi, \tag{V}$$

That is: if φ is true then it is possible (for someone) to know that φ is true. This might seem like a reasonable thesis, and certainly not enough by itself to unseat verificationism as a tenable metaphysical stance.

We say that an agent who knows every truth is *omniscient*. The following theorem, appears in Fitch (1963), where it is attributed to an anonymous referee of an earlier (1945) unpublished paper:[3]

[3] The referee, it later transpired, was Alonzo Church (Brogaard and Salerno 2008).

Theorem 7 (Fitch) *For each agent who is not omniscient, there is a true proposition which that agent cannot know.*

The contrapositive of this statement is that if an agent *can* know every true proposition, then that agent is omniscient, i.e. *does in fact know* every true proposition. In a logical notation with quantifiers over propositions, we would write this as (F).

$$(\forall\varphi(\varphi \rightarrow \Diamond K_i\varphi)) \rightarrow \forall\varphi(\varphi \rightarrow K_i\varphi) \tag{F}$$

Given (F), the verificationist thesis (V) entails that the agent is omniscient! This theorem of Fitch has been used to argue for metaphysical realism (Hart and McGinn 1976), or at least against verificationism, and some verificationists have even seen it as a serious challenge worth addressing (Dummett 1976).

In formalising philosophical arguments, and drawing philosophical conclusions from the formal conclusions, ideally one would not add or lose anything from the original philosophical statements. However, in useful formalisations some things are of course lost or added, for example Fitch himself noticed that the *temporal* aspect is missing from his rather limited analysis of the "value concept" (his expression) of knowledge:

> "For purposes of simplification, the element of time will be ignored in dealing with these various concepts." (Fitch 1963, p. 136)

Since dynamic epistemic logic deals with the interface between time and knowledge, it is unsurprising that DEL-style thinking has led to a resolution of this paradoxical result. van Benthem (2004) observes that essential to Fitch's result (and indeed to others from the same paper) is a judicious manipulation of a Moore-like sentence $p \wedge K_i p$.

The DEL methodology is a correct way to integrate the "element of time" into the model. Recall the example given in Fig. 1. There we saw an annoucement that was such that *after* the announcement *the announced sentence is not known* (indeed, its negation is known). We mentioned that the intuitive idea is that one learns, when a Moore sentence is announced, is that is *was* true. Blindspots like Moore sentences *cannot*, as a matter of principle, be known by any agent ideal enough to have positive introspection (i.e. such that if she knows something then she knows that she knows it). The existence of unknowable truths is indeed the opposite of verificationism, but this kind of unknowable truth is surely not what the verificationist has in mind when denying their existence! Philosophers could long have suspected that there was something fishy going on with Fitch's paradox, and the DEL methodology indeed reveals precisely what this is.

Building on the observations of van Benthem, Balbiani et al. (2008) flesh out the $\Diamond$ operator, characterising 'knowability' as 'known after some announcement'. They therefore develop 'arbitrary public announcement logic', in which $\Diamond$, written as $\langle ! \rangle$, means 'after some public announcement …', and give expressivity and completeness results for the resulting logic.

3 Belief change

One area in which DEL as presented is inadequate is in dealing with *beliefs* rather than knowledge. Beliefs have the characteristic that they must be *revisable*: you might believe p right now, but could learn that $\neg p$. The form of DEL that we have presented is not appropriate for modeling this kind of reasoning. Yet belief revision is a logical process which should have rational constraints to it. It is also a process that interactive epistemologists from game theory are recognising as important, in what Brandenburger (2007) calls the "belief-based approach". We therefore now turn to look at ways in which the DEL methodology can be cashed out to reason properly about *beliefs* rather than about knowledge.

The dynamic perspective characterises knowledge as the strong, indefeasible, result of "hard information". In many situations the relevant hard information may be lacking, with enough uncertainty remaining at the epistemic level for the agent to have recourse to *beliefs* as a basis for action. In contrast with knowledge, the beliefs of an agent are defeasible, so the agent is ready to *give up* her beliefs in the light of incoming information. This opens the door to so-called "*belief revision*" theory.

3.1 AGM

Initial logical investigations in belief revision were led by Alchourrón et al. (1985) (the authors called belief revision "theory change"). In the framework of that paper, a belief state is represented by a set Γ of formulas (of e.g. a propositional language, that we call the "object language"). Revision is then a syntactic operation taking a set Γ and a formula φ, and returning $\Gamma * \varphi$, a new set of sentences. Alchourrón et al. (1985) proposed and studied some *rational constraints* on such functions, known as the AGM postulates. As proved in Grove (1988), these constraints have an elegant semantic analogue: a revision function $\Gamma * (\cdot)$ respects the postulates just if it is definable in terms of a pre-order $\leq_\Gamma$ over the maximally consistent sets of sentences of the object language, with $\Gamma * \varphi \vdash \psi$ just if ψ holds at *all* $\leq_\Gamma$-*minimal* φ-states. It is this semantic version of the AGM postulates that form the basis of the work we will now summarise.

3.2 Belief dynamics

In the AGM paradigm, only one agent is considered, and the logical languages considered cannot refer to the beliefs of the agent. The epistemic approach based on equivalence relations can be extended by adding pre-orders to the models, giving us a natural way of representing multi-agent belief revision. These models are proposed and discussed in Board (2002), van Benthem (2007a), Baltag and Smets (2008).

Definition 8 An epistemic-plausibility model for the agents N is a tuple $(W, (\sim_i, \leq_i)_{i\in N}, V)$, where$(W, (\sim_i)_{i\in N}, V)$ is an epistemic model and each $\leq_i$ is a well-founded pre-order (reflexive and transitive relation) over W.

Definition 9 *(A priori/ a posteriori) Most plausible elements*

- For all $X \subseteq W$, let $\beta_i(X) = min_{\leq_i}(X) = \{w : w \text{ is } \leq_i \text{-minimal in } X\}$.
- For all $w \in W$, let $\mathcal{B}_i[w] = \beta_i(\mathcal{K}_i[w])$.

We write $w \rhd_i^{\mathcal{B}} v$ iff $v \in \mathcal{B}_i[w]$, and $w \rightarrow_i^X v$ iff $v \in \beta_i(\mathcal{K}_i[w] \cap X)$.

As in epistemic models, where *common knowledge* is an important concept, here we will want to define *common belief*.

Definition 10 (*Common belief*) For each $G \subseteq I$, let $\rhd_G^*$ be the reflexive-transitive closure of $\bigcup_{i \in G} \rhd_i^{\mathcal{B}}$. $[w]_G^* = \{w' \in W \mid w \sim_G^* w'\}$.

3.3 Doxastic-epistemic logic

The logical language used in Baltag and Smets (2008) to describe epistemic-plausibility models is a propositional modal language with three families of modal operators, which we extend here with "common belief" operators.

Definition 11 (*Epistemic Doxastic Language*) The language $\mathcal{L}_{EDL}$ is defined as follows:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid B_i^{\varphi}\varphi \mid CB_G\varphi,$$

where i ranges over N, p over a countable set of proposition letters PROP and $\emptyset \neq G \subseteq I$.

We write $\bot$ for $p \wedge \neg p$ and $\top$ for $\neg\bot$. A formula $K_i\varphi$ should be read as "i knows that φ", $CB_G\varphi$ as "it is common belief among group G that φ." The formula $B_i^{\varphi}\psi$, should be read " conditional on φ, i believes that ψ." These formulas are interpreted in epistemic plausibility models as follows:

Definition 12 (*Truth definition*) We write $||\varphi||^{\mathcal{M}}$ for $\{w \in |\mathcal{M}| : \mathcal{M}, w \Vdash \varphi\}$. We omit $\mathcal{M}$ when it is clear from the context.

$$\begin{array}{ll} \mathcal{M}, w \Vdash B_i^{\psi}\varphi & \text{iff } \forall v(\text{if } w \rightarrow_i^{||\psi||^{\mathcal{M}}} v \text{ then } \mathcal{M}, v \Vdash \varphi) \\ \mathcal{M}, w \Vdash CB_G\varphi & \text{iff } \forall v(\text{if } w \rhd_G^* v \text{ then } \mathcal{M}, v \Vdash \varphi) \end{array}$$

Simple belief conditional only on i's information at a state w is definable using the conditional belief operator: $B_i\varphi = B_i^{\top}\varphi$, since: $\mathcal{M}, w \Vdash B_i^{\top}\varphi$ iff $\forall v(\text{if } w \rhd_i^{\mathcal{B}} v \text{ then } \mathcal{M}, v \Vdash \varphi)$.

The advantages of the dynamic approach over the static syntactic approach of AGM are threefold. Firstly, we find as in the case of PAL, that there is a solution to the problem of Moore sentences. That is, the Success Postulate $\Gamma \star \varphi \vdash \varphi$ fails in a contained fashion, notably for blindspots. By dealing with epistemic states[4], we also have a natural way of trearing *iterated revision*. Finally, the approach copes seemlessly with the case of multiple agents.

[4] i.e. including "pre-encoding" the information about how an agent will change her beliefs cf. van Benthem (2007a).

4 Game theory

A strategic game is an interactive decision process in which are specified the different choices of action that the players have, and their preferences over the outcomes of those (collective) choices. Players may have different first-order information and beliefs about the state of the world but also, and just as importantly, they may have different higher-order information and beliefs, i.e. information and beliefs about what other agents know/believe.

Now assume that two players are to play a game for the first time together. Assume that their preferences over outcomes are common knowledge between them. What would be a 'good' or 'correct' decision? One possible answer is that they should choose an action that survives so called 'iterated elimination of strictly dominated strategies' (IESDS Osborne and Rubinstein 1994).

Theorem 13 (Tan and Werlang 1988) *A choice c can rationally be chosen under common true belief of rationality iff c survives IESDS.*

This use of a decision-theoretic approach shifts the focus from the usual equilibrium notions, i.e. Nash equilibrium and its refinements, to game reductions, and is part of the *epistemic program* in game theory.

Equilibria re-emerge, however, in the form of fixpoints. Firstly, common knowledge is itself a fixpoint, and secondly the iterative processes themselves implicitly involve analyses in terms of fixpoints. This is revealed in two different ways in Apt and Zvesper (2007), van Benthem (2007b). Apt and Zvesper (2007) provides a syntax in terms of modal fixpoint languages for reasoning about rationality and elimination of dominated strategies. van Benthem (2007b) takes a new and more dynamic perspective, analysing the algorithm of iteration itself as a series of public announcements that the relevant strategies will not be played (because of the rationality of the players involved). This leads to a partial answer to the question *how* the epistemic conditions, like common belief of rationality, can come about. An important point here is that after an announcement that players are rational it need not necessarily be the case that the players are rational (notice the similarity with the case of Moore sentences). However, the announcements will eventually stabilise, leading to the appropriate kind of equilibrium.

More generally, games provide a good setting and test-bed for the merging of logics of action, belief and preference. We now consider two case studies to illustrate the interplay between (dynamic) logic and games.

5 Agreeing to disagree

In the setting we have introduced to encode the information and the beliefs of agents we could say that the pre-order encodes the *prior beliefs* of agents about the state of the world and that the cells of the epistemic partition encode the *hard, private information* received by each agent. We have seen that in game theory higher-order beliefs have a decisive status. Other areas where they are crucial include e.g. financial economics (roughly, can agents make use of private information when all agents value the

'goods' in the same way) and theories of competitive equilibrium under uncertainty. A natural question (solved by Aumann (1976) for a natural probabilistic setting) can be raised about epistemic plausibility models: namely, can agents agree to disagree? More precisely, is it possible for it to be common knowledge between two agents with the same prior beliefs that they have different beliefs about some facts of the world?

For the probablistic Aumann proved that this is impossible.

Theorem 14 (Aumann 1976) *If two agents 1 and 2 have the same prior, then if it the posterior probabilities they assigned to A are common knowledge between them at ω then these posterior probabilities are equal.*

Recently Dégremont and Roy (2009) proved that this is true also in the setting of epistemic plausibility models in which the plausibility ordering is well-founded. In fact they proved something stronger, namely that if it is common *belief* between two agents that they have different (posterior) beliefs about some fact of the world, then they have different priors.

Theorem 15 (Agreement theorem—Common Belief) *If a well-founded epistemic plausibility model $\mathcal{M}$ satisfies $\mathcal{M}, w \Vdash C B_{\{i,j\}}(B_i p \wedge \neg B_j p)$ for some $w \in W$, then i and j have different priors in $\mathcal{M}$.*

It is easy to see that this result implies that common knowledge of disagreement is impossible in a *well-founded* epistemic plausibility model. Interestingly Dégremont and Roy (2009) show that when well-foundedness is weakened to local-wellfoundeness (the plausibility ordering of every cell of the epistemic partition is well-founded) the agreement theorem fails. The preceding theorem is a semantic result about epistemic plausibility models. Doxastic logics are naturally interpreted on epistemic plausibility models (cf. Subsection 3.3 for an example of such a doxastic logic). Does it mean that agreement theorems can be considered as theorems of some doxastic logic? The answer is yes, but the logic has to be sufficiently expressive. In particular the basic doxastic logic with common belief cannot define common prior.

Fact 16 *The class of epistemic plausibility frames that satisfies common prior is not definable in $\mathcal{L}_{EDL}$.*

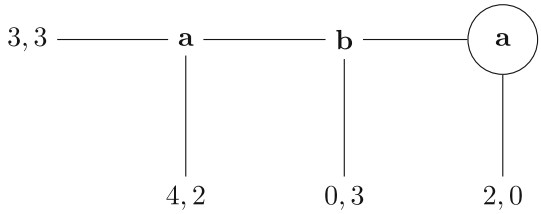
Dégremont and Roy (2009) give a (finite) formal axiomatic derivation in some hybrid logic. However the validities of this logic itself are not recursively enumerable, ruling out the possibility of giving a finite complete axiomatization of its validities.

Finally the paper addresses the issue of whether disagreeing agents with the same plausibility ordering (same prior) will reach agreement by iteratively communicating their beliefs for a sufficiently long time. It is shown that in particular in the case of communicating through public announcements agreement will be reached.

6 Backward induction

Extensive games, sometimes called "dynamic games", also lend themselves naturally to analysis in terms of belief dynamics. An extensive game is a tree, with each non-terminal node v being a decision point for one player $\rho(v)$, and each terminal node o

Fig. 5 An extensive game



being an "outcome", over which the players have preferences in the same way as in a strategic (or "static") game: $o <_i o'$ means that player i strictly prefers the outcome o' over o. The difference between static and dynamic games is that in the latter, players could in principle *change* their beliefs *as the game is played*. So this is naturally an area where logics of belief change can be applied.

In (finite) extensive-form games of perfect information, a standard solution concept is 'backward induction', which has the pleasant feature that for "generic" games (in which each player's preference relation is strict), it yields a unique solution. For a formal definition of backward induction see Osborne and Rubinstein (1994).

Aumann (1995) provided an epistemic foundation for backward induction that has since been criticised, notably by Stalnaker (1996).[5] What Aumann showed is that common knowledge of "substantive rationality" entails the backward induction outcome. Substantive rationality means rationality, *at every node*. Halpern (2001) argues persuasively that ambiguities in the way "substantive rationality" can be understood explain the difference between Aumann and Stalnaker's views. To understand the objections raised to Aumann's putative epistemic foundation, it can help to consider the simple game in Fig. 5. Inuitively speaking, Aumann's argument looks like this: Assume common knowledge of (substantive) rationality. Then in particular, a is rational at her last decision node, and so will play down, since $4 > 3$. Since all this is common knowledge, b knows that a will go down, and so he will play down, since $3 > 2$. Then a knows all of the above, so knows she has a choice between 0 and 2; clearly rationality means that she will choose 2, i.e. play down. All seems well, but notice now that it cannot have been possible to have (common knowledge of) substantive rationality at a's last decision node, since the argument just given proves that *reaching that decision node contradicts common knowledge of substantive rationality*!

Nonetheless, backward induction *does* make sense, and there surely must be some condition on their beliefs that will mean players arrive at the backward induction outcome. What was not explicitly present in Aumann's analysis was the idea that players must reason that *if they were* to reach such-and-such node, then they would all still be rational. The reasoning behind the backward induction algorithm therefore invokes some notion of counterfactual belief. This idea has been explored in the game-theoretical literature, and a number of ad-hoc frameworks have been proposed along these lines (for example by Samet (1996), Arieli (2008)), but from the way we have pre-

[5] A great deal of literature exists on the epistemic foundations of backward induction, going back before these two papers; we do not attempt any kind of survey here, but Bicchieri (1989) and Binmore (1987) are important early contributions.

sented backward induction reasoning, it should be clear that doxastic-epistemic logic will help give a formal analysis.

And indeed an analysis in terms of doxastic-epistemic logic *with public announcement operators* allows us to cash out these intuitions in a straighforward formal framework. Baltag et al. (2009) define a logical language with announcement operators $[!\varphi]$ and $[!]$ (for "arbitrary announcements", see Sect. 2), and prove that common knowledge of "stable belief" in "dynamic rationality" entails common belief in the backward induction outcome.[6] "Dynamic rationality" means forward-looking rationality, i.e. the past history is entirely ignored. Stable belief is belief that is invariant under acquired information: that is, no matter what is learned (announced), the belief is maintained. Formally, stable belief is an operator defined by composing the arbitrary announcement and belief operators: $[!]B$.

The authors also introduce *stable true belief*, $Stb_i\varphi$, defined as $K_i[!]B_i\varphi \wedge [!]\varphi$, and show that the weaker and simpler condition of common stable true belief of rationality is sufficient to entail the backward induction outcome. Stable true belief is an interesting epistemic notion in its own right: it is positively introspective, but not negatively introspective, and indeed when applied to ontic facts (basic propositions p), it coincides with what Stalnaker (2006) calls "knowledge". (There is divergence when it comes to non-ontic facts, i.e. facts with some epistemic content, that might *change* when some announcement is made.)

We can think of the condition of stable belief as partially constraining the players' "belief revision policies": it is a sort of 'optimism', according to which irrespective of what true information they receive about what happens in the game, including perhaps information that contradicts rationality, they will maintain the belief that players will be rational.

7 Further directions

The interplay between game theory and logic represented by these two case studies continues a tradition instantiated by a number of other authors. For example the work of de Bruin (2004), who provides a purely syntactic approach to the analysis of solution concepts in a formal logical language, is in this tradition, as is that of Pauly (2001), who analyses coalitional games from a logical perspective.

Knowledge and its dynamics are also very important in computer science, as has been recognised by research since the 80s, documented in Fagin et al. (1995).

Some recent work (Hendricks 2001; Dégremont and Gierasimczuk 2009) considers the interface between DEL and formal learning theory. By analysing the temporal doxastic structure underlying formal learning theory, this approach provides additional insight into the semantics of inductive learning. By importing the ideas, problems and methodology from Learning Theory, logics of epistemic and doxastic change get enriched by new (inductive) learning scenarios, new concepts and new problematic perspectives.

[6] And if rationality is added as a condition then the players will actually play that backward induction outcome.

Other issues with philosophical significance include the phenomena of intention and preference change. Girard (2008) extensively analyses logics of preferences, in particular presenting a logic of 'ceteris paribus' preference, where 'ceteris paribus' is given the reading 'everything else being equal'. Ceteris paribus logic can also be applied to interpretations of the modality other than preference. Liu (2008) considers dynamic logics of preference change and Roy (2008) formalises different notions of *intention* and examines also their dynamics. Roy argues that intention can be used to explain the phenomenon of coordination in so-called 'Hi-Lo' games, i.e., coordination games in which there exists a Pareto-dominant Nash equilibrium.

This is a burgeoning area in which there are many further topics to be explored. We anticipate that in the future interaction-oriented logics will embrace the problem of belief *merge*, as well as put in evidence and analyze the epistemic aspects of cooperative game theory and social choice theory.

The dynamic-epistemic project internalizes information flow, belief change, inductive or strategic reasoning into the formal language, bringing the logical analysis of the informational dimension of interaction one step further. In this paper we have sketched some applications of DEL that illustrate how it constitutes a natural foundation for a logical theory of rational and intelligent interaction.

References

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic, 50*(2), 510–530.

Apt, K. R., & Zvesper, J. A. (2007). *Common beliefs and public announcements in strategic games with arbitrary strategy sets*. Under review. Available from http://arxiv.org/pdf/0710.3536v2.

Arieli, I. (2008). *Backward induction and common strong belief of rationality*. In K. R. Apt & R. van Rooij (Eds.), *New perspectives on games and interaction*, volume 4 of *Texts in logic and games* (pp. 265–282).

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics, 4*(6), 1236–1239.

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior, 8*, 6–19.

Balbiani, P., Baltag, A., van Ditmarsch, H., Herzig, A., Hoshi, T., & de Lima, T. (2008). 'Knowable' as 'known after an announcement'. *The Review of Symbolic Logic, 1*(3), 305–334.

Baltag, A., & Moss, L. S. (2004). Logics for epistemic programs. *Synthese, 139*, 165–224.

Baltag, A., & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, & M. Wooldridge (Eds.), *Logic and the foundations of game and decision theory (LOFT 7)*, volume 3 of *Texts in logic and games* (pp. 9–58). Amsterdam University Press.

Baltag, A., Moss, L. S., & Solecki, S. (1999). *The logic of public announcements, common knowledge and private suspicions*. Technical Report SEN-R9922, Centrum voor Wiskunde en Informatica.

Baltag, A., Smets, S., & Zvesper, J. A. (2009). Keep 'hoping' for rationality: A solution to the backward induction paradox. *Synthese, 169*(2), 301–333.

Bicchieri, C. (1989). Self-refuting theories of strategic interaction: A paradox of common knowledge. *Erkenntnis, 30*, 69–85.

Binmore, K. (1987). Modeling rational players, part I. *Economics and Philosophy, 3*, 179–214.

Blackburn, P., de Rijke, M., & de Venema, Y. (2001). *Modal logic*. Cambridge, UK: Cambridge University Press.

Board, O. (2002). Dynamic interactive epistemology. *Games and Economic Behavior, 49*, 49–80.
Brandenburger, A. (2007). The power of paradox: Some recent developments in interactive epistemology. *International Journal of Game Theory, 35*(4), 465–492.
Brogaard, B., & Salerno, J. (2008). Fitch's paradox of knowability. In Zalta, E. N. (Ed.), *Stanford Encyclopedia of Philosophy.*
de Bruin, B. (2004). *Explaining games: On the logic of game theoretic explanations*. PhD thesis, ILLC, Amsterdam.
Dégremont, C., & Gierasimczuk, N. (2009). Can doxastic agents learn? On the temporal structure of learning. In X. He, J. F. Horty, & E. Pacuit, (Eds.), *LORI*, volume 5834 of *Lecture Notes in Computer Science* (pp. 90–104). Springer.
Dégremont, C., & Roy, O. (2009). *Agreement theorems in dynamic-epistemic logic*. In A. Heifetz (Ed.) *TARK '09* (pp. 91–98). New York, NY, USA: ACM.
Dummett, M. (1976). What is a theory of meaning? (II). In G. Evans & J. McDowell (Eds.), *Truth and meaning, chapter 4.* (pp. 67–137). Oxford: Clarendon Press.
Fagin, R., Halpern, J. Y., Vardi, M., & Moses, Y. (1995). *Reasoning about knowledge*. Cambridge, MA, USA: MIT Press.
Fitch, F. B. (1963). A logical analysis of some value concepts. *Journal of Symbolic Logic, 28*(2), 135–142.
Gerbrandy, J. (1999). *Bisimulations on planet Kripke*. PhD thesis, ILLC, Amsterdam.
Gerbrandy, J., & Groeneveld, W. (1997). Reasoning about information change. *JoLLI, 6*, 147–169.
Girard, P. (2008). *Modal logic for belief and preference change*. PhD thesis, University of Stanford, 2008. ILLC Dissertation Series DS-2008-04.
Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic, 17*(2), 157–170.
Halpern, J. Y. (2001). Substantive rationality and backward induction. *Games and Economic Behavior, 37*, 425–435.
Hart, W. D., & McGinn, C. (1976). Knowledge and necessity. *Journal of Philosophical Logic, 5*, 205–208.
Hendricks, V. F. (2001). *The convergence of scientific knowledge: A view from the limit. Studia Logica Library Series: Trends in Logic*. Dordrect: Kluwer.
Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca: Cornell University Press.
Liu, F. (2008). *Changing for the better: Preference dynamics and agent diversity*. PhD thesis, University of Amsterdam.
Moore, G. E. (1942). A reply to my critics. In P. A. Schilpp, (Ed.), *The philosophy of G.E. Moore*, volume 4 of *The library of living philosophers*, pages 535–677. Northwestern University, Evanston, IL.
Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge: MIT Press.
Pauly, M. (2001). *Logic for social software*. PhD thesis, University of Amsterdam. ILLC Dissertation Series DS-2001-10.
Plaza, J. A. (1989). Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, & Z. W. Ras (Eds.) *Proceedings of the 4th international symposium on methodologies for intelligent systems* (pp. 201–216).
Roy, O. (2008). *Thinking before acting: Intentions, logic, rational choice*. PhD thesis, ILLC, Amsterdam.
Samet, D. (1996). Hypothetical knowledge and games with perfect information. *Games and Economic Behavior, 17*, 230–251.
Sorensen, R. A. (1988). *Blindspots*. New York: OUP.
Stalnaker, R. C. (1996). Knowledge, beliefs and counterfactual reasoning in games. *Economics and Philosophy, 12*, 133–163.
Stalnaker, R. C. (2006). On logics of knowledge and belief. *Philosophical Studies, 128*, 169–199.
Tan, T. C.-C., & da Costa Werlang, S. R. (1988). The Bayesian foundations of solution concepts of games. *Journal of Economic Theory, 45*(2), 370–391.
van Benthem, J., van Jan, E., & Kooi, B. P. (2006). Logics of communication and change. *Information and Computation, 204*(11), 1620–1662.
van Benthem, J. (1996). *Exploring logical dynamics*. Stanford, CA: CSLI.
van Benthem, J. (2003). One is a lonely number. *ILLC Prepublication*, PP-2003(07).
van Benthem, J. (2004). What one may come to know. *Analysis, 64*(2), 95–105.
van Benthem, J. (2007). Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics, 17*(2), 129–155.

van Benthem, J. (2007). Rational dynamics and epistemic logic in games. *International Game Theory Review, 9*(1):13–45, (Erratum reprint, 9(2), 377–409).

van Ditmarsch, H. (2000). *Knowledge games*. PhD thesis, Universiteit van Amsterdam.

van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2007). *Dynamic epistemic logic*, volume 337 of *Synthese Library Series*. Springer.